THE STANDARD ERRORS OF VARIOUS TEST STATISTICS WHEN THE TEST ITEMS ARE SAMPLED

A Technical Report

prepared by

FREDERIC M. LORD

Office of Naval Research Contract Nonr-694(00)
Project Designation NR 151-113

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY

December, 1953

THE STANDARD ERRORS OF VARIOUS TEST STATISTICS WHEN THE TEST ITEMS ARE SAMPLED

A Technical Report

prepared by

FREDERIC M. LORD

Office of Naval Research Contract Nonr-694(00)

Project Designation NR 151-113

EDUCATIONAL TESTING SERVICE

Princeton, New Jersey

December, 1953

THE STANDARD ERRORS OF VARIOUS TEST STATISTICS

WHEN THE TEST ITEMS ARE SAMPLED

Frederic M. Lord

Abstract

Suppose that a large number of forms of the same test are administered to the same group of examinees, each form consisting of a random sample of items drawn from a common pool of items. If some test statistic is computed separately for each form of the test, the value obtained will (ignoring practice effect, fatigue, etc.) differ from form to form because of sampling fluctuations. The standard deviation of the values obtained represents, approximately, the standard error of the test statistic when the test items are sampled.

Formulas for such standard errors are here derived for a) the test score of a single examinee, b) the mean test score of a group of examinees, c) the standard deviation of the scores of the group, d) the Kuder-Richardson reliability of the test, formula 20, e) the Kuder-Richardson reliability, formula 21, f) the test validity. In large samples, the foregoing statistics (with the possible exception of d) are approximately normally distributed, so that significance tests can be made by familiar procedures.

Consideration is given to the relation of certain of the foregoing standard errors to the conventional standard error of measurement, to the Kuder-Richardson reliability coefficients 20 and 21, and to the Wilks-Votaw criterion for parallel tests. Practical applications of the results are briefly discussed. In particular, it is concluded that the Kuder-Richardson formula-21 reliability coefficient should properly be used in certain practical situations instead of the commonly preferred formula-20 coefficient.

THE STANDARD ERRORS OF VARIOUS TEST STATISTICS WHEN THE TEST ITEMS ARE SAMPLED*

Frederic M. Lord

Suppose that the same test is administered to a large number of separate groups of examinees, the groups being random samples all drawn from the same population; and suppose that some test statistic is computed separately for each sample of examinees. The value obtained for this test statistic will, of course, differ from sample to sample because of sampling fluctuations. The standard deviation of these values over a very large number of samples is the standard error of the test statistic *when examinees are sampled*. For convenience, this type of sampling will be referred to as *type 1* sampling.

On the other hand, suppose that a large number of forms of the same test are administered to the same group of examinees, each form consisting of a random sample of items drawn from a common population of items; and suppose that some test statistic is computed separately for each form of the test. Let us assume for theoretical purposes that the examinees do not change in any way during the course of testing, i.e., that there is no practice effect, no fatigue, etc. The value computed for the test statistic will still, of course, differ from form to form because of sampling fluctuations. The standard deviation of these values over a very large number of samples is the standard error of the test statistic *when the test items are sampled*. This type of sampling will be referred to as a *type 2* sampling. Test forms constructed by type 2 sampling will be called *randomly parallel forms* or *randomly parallel tests*.

*The writer is indebted to Professor S. S. Wilks who has checked over certain critical portions of a draft of this paper.

Type 1 standard error formulas have long been available and are sometimes incorrectly used in situations where sampling of test items is of crucial importance. The present paper is concerned with deriving formulas for the type 2 standard errors of certain test statistics. Formulas for the two kinds of standard errors may usually be readily distinguished on a superficial level by the following characteristics, which underscore the essential difference between them: type 1 standard errors are usually obviously proportional to some power (positive or negative) of the number of examinees in the sample -- most commonly inversely proportional to the square root of this number -- and are usually much less obviously and simply related, if at all, to the number of items in the test; type 2 standard errors have the corresponding characteristic with respect to n , the number of items in the sample.

Notation and Summary of Formulas

The test statistics with which the present study is concerned are primarily the following:

t_a -- the observed test score of examinee a , obtained by counting the number of items answered correctly on a single test.

$\bar{t}$ -- the mean of the scores obtained by the N examinees on a single test. $\bar{t} = \sum_a t_a/N$.

s_t -- the standard deviation of the scores obtained by the N examinees on a single test. $s_t^2 = \sum_a t_a^2/N - \bar{t}^2$.

r_{21} -- the Kuder-Richardson reliability coefficient, formula 21.

$$r_{21} = \frac{n}{n-1}\left[1 - \bar{t}(n - \bar{t})/ns_t^2\right] \quad .$$

r or r_{20} -- the Kuder-Richardson reliability coefficient, formula 20. $r = \frac{n}{n-1}\,(1 - \sum_i s_i^2/s_t^2)$ (symbols explained in the succeeding list).

r_{ct} -- the correlation of the test score with any external variable, c . $r_{ct} = s_{ct}/s_c s_t$.

Considerable care in defining notation must be taken here in order to avoid serious confusion. Additional symbols that will be used are listed below for easy reference.

x_{ia} -- the "score" of examinee a on item i . $x_{ia} = 1$ if the item is answered correctly. $x_{ia} = 0$ otherwise.

n -- the number of items in a single form of a test, i.e., in a single sample. The subscript i runs from 1 to n .

N -- the number of examinees in a single group of examinees. The subscript a runs from 1 to N .

m -- the number of items in a finite population of items.

p_i -- the observed "difficulty" of item i for the N examinees tested. $p_i = \sum_a x_{ia}/N$.

$q_i = 1 - p_i$.

z_a -- the "proportion-correct score" of examinee a ; the proportion of the items in a single test answered correctly by examinee a . $z_a = t_a/n$.

$\bar{z}$, $\bar{c}$, etc. -- the mean of the N values of z , c , etc. $\bar{z} = \sum_a z_a/N$, etc.

M(p) -- the mean of the n observed values of p_i for the n items in the test administered. $M(p) = \sum_i p_i/n$.

s_c , s_z , etc. -- the standard deviation of the N values of c , z , etc. $s_z^2 = \sum_a z_a^2/N - \bar{z}^2$, etc.

s_i -- the standard deviation of x_{ia} for fixed i .

$$s_i^2 = \sum_a x_{ia}^2/N - (\sum_a x_{ia}/N)^2 = p_i q_i \quad .$$

s_{ct} , etc. -- the covariance (over examinees) of c and t , etc. $s_{ct} = s_c s_t r_{ct} = \sum_a (c_a - \bar{c})(t_a - \bar{t})/N$.

s_{ic} , s_{iz} , s_{it} -- the covariance (over examinees) of c_a , z_a , or t_a , respectively, with x_{ia} , for fixed i .

$$s_{it} = s_i s_t r_{it} = \sum_a (x_{ia} - p_i)(t_a - \bar{t})/N \quad .$$

s(p) -- the standard deviation of the n observed values of p_i for the n items in the test administered.

$$s^2(p) = \sum_i p_i^2/n - M^2(p) \quad .$$

$s(s_{iz})$, $s(s_{it})$, etc. -- the standard deviation of the n observed values of s_{iz} , s_{it} , etc. for the n items in the test administered. $s^2(s_{it}) = \sum_i s^2_{it}/n - (\sum_i s_{it}/n)^2$.

$s(s_{ic},s_{it})$ -- the covariance (over items) of s_{ic} and s_{it} . $s(s_{ic},s_{it}) = \sum_i s_{ic}s_{it}/n - (\sum_i s_{ic}/n)(\sum_i s_{it}/n)$.

r_{ic} , r_{it} , r_{iz} -- the correlation of c_a , t_a , or z_a , respectively with x_{ia} , for fixed i . $r_{it} = s_{it}/s_i s_t$

It should be noted that all the statistics in the foregoing list are observed sample statistics relating to a given sample. There are two kinds of statistics listed, typified, in the simplest case, by $\bar{z} = \sum_a z_a/N$ and $M(p) = \sum_i p_i/n$. Population parameters have not been listed but will be designated, when needed, by the use of Greek letters. The following additional symbols, relating to the totality of all possible samples of test items (type 2 sampling), will be used.

$E(x)$ -- the <u>expected value</u> of x ; the arithmetic mean of the statistic x over all possible samples.

$S.E.(x)$ -- the standard error of the statistic x ; the standard deviation of the statistic x over all possible samples. $S.E.^2(x) = E(x^2) - [E(x)]^2$.

Var x -- the <u>sampling variance</u>. Var $x = S.E.^2(x)$.

$Cov(x,y)$ -- the <u>sampling covariance</u> of the statistics x and y over all possible samples. $Cov(x,y) = E(xy) - E(x)E(y)$.

Table 1 summarizes the more important of the type 2 standard errors derived in the present paper. For purposes of comparison, the last column of the table, when appropriate, gives the corresponding usual type 1 formulas for the standard error for the case where the test scores are assumed to be normally distributed. The standard error formulas in both columns are large-sample formulas, in general, and observable sample statistics have been substituted for the corresponding population values throughout.

At a later point it will be proven that in large samples of the second type all of the test statistics in the left-hand column of Table 1, with the possible exception of the Kuder-Richardson formula 20 reliability coefficient, have an asymptotically normal sampling distribution.

Table 1

Standard Errors of Test Statistics

Statistic	Type 2 (When Items Are Sampled)	Type 1 (When Examinees Are Sampled)
t_a	$\sqrt{\frac{t_a(n - t_a)}{n}}$	--
$\bar{t}$	$\sqrt{n}\ s(p)$	$\frac{s_t}{\sqrt{N}}$
s_t	$\frac{\sqrt{n}\ s(s_{it})}{s_t}$	$\frac{s_t}{\sqrt{2N}}$
r_{20}	$\frac{\sqrt{n}}{s_t^2} \sqrt{s^2(s_i^2) - 4(1 - r_{20}) s(s_i^2, s_{it}) + 4(1 - r_{20})^2 s^2(s_{it})}$	*
r_{21}	$\frac{1}{\sqrt{n}\ s_t^2} \sqrt{(n - 2\bar{t})^2 s^2(p) + 4n^2(1 - r_{21})^2 s^2(s_{it}) - 4n(1 - r_{21})(n - 2\bar{t}) s(p_i, s_{it})}$	*
r_{ct}	$\frac{\sqrt{n}}{s_t} \sqrt{\frac{1}{s_c^2} s^2(s_{ic}) - \frac{2r_{ct}}{s_c s_t} s(s_{ic}, s_{it}) + \frac{r_{ct}^2 s^2(s_{it})}{s_t^2}}$	$\frac{1 - r_{ct}^2}{\sqrt{N}}$

*Not known to writer.

Illustrative Examples and Discussion of the Standard Errors

Suppose that Form A of a certain 135-item test has been administered. Several parallel forms of this same test are to be administered in the future. Each form is administered to a different group of examinees. The groups of examinees may be considered as random samples drawn from the same population. Each group is so large that differences between groups due to sampling of examinees may be ignored.* It is found that the mean, standard deviation, and Kuder-Richardson formula 20 reliability of the scores on Form A are 63.5, 21.5, and 0.95, respectively. How much may we expect the means to vary from form to form?

The required value of $s(p)$ can be determined directly from item analysis data; or it can be calculated from the three numerical values given by solving for $s^2(p)$ Tucker's modification (8) of the equation for the Kuder-Richardson formula 20 reliability, the result being:

$$s^2(p) = \frac{s_t^2}{n}\left(\frac{n-1}{n} r_{20} - 1\right) + \frac{\bar{t}}{n} - \frac{\bar{t}^2}{n^2} \tag{1}$$

We find that $s^2(p) = .0538$.

The large-sample estimate of the type 2 standard error of the mean is found to be $S.E._2(\bar{t}) = 2.7$. (The subscript "2" is used here, and the subscript "1" is used below, to indicate type 2 and type 1 standard errors, respectively. Hereafter, type 2 sampling will be understood, unless otherwise specifically indicated.) If the same test were admin-

*Useful formulas for dealing simultaneously with sampling of items and sampling of examinees have been developed by the writer for certain of the statistics studied here. Some such formulas are recently independently reported in Hooke, R., "Sampling from a matrix, with applications to the theory of testing." Princeton University Statistical Research Group, Memorandum Report 53, 1953. (Dittoed.)

istered to random groups of 135 examinees, the type 1 standard error would be $S.E._1(\bar{t}) = 1.8$.

On the basis of the foregoing, we may expect that parallel forms of the test would not differ from each other in mean score by as much as $2\sqrt{2}S.E._2(\bar{t}) = 7.6$ points more than one time in twenty. If the parallel forms are carefully constructed by matching items from form to form on difficulty and item-test correlation rather than by random sampling of items, it may well be that the forms will not differ from each other as much as the foregoing formulas would indicate. On the other hand, it is not unlikely that supposedly parallel forms of a test may, because of the unconscious bias of the test constructor, often be found in fact to be less parallel than would be expected if each form were a random sample of test items.

In many kinds of statistical experiments it is commonly not merely desirable but actually necessary to select cases by random sampling rather than by stratified sampling, even though random sampling gives rise to larger sampling fluctuations. The reason is, first, that random sampling tends to avoid unintentional bias; and, second, that the standard errors arising from random sampling are known and easily used, whereas those arising from stratified sampling are often either unknown or excessively cumbersome to use. Similarly, and for the same reasons, it will be desirable in certain kinds of experimental work, to use parallel forms composed of items selected at random rather than in any other way.

Suppose, for example, it is desired to investigate the relation of length of reading passage to validity in a reading comprehension test. The experimenter might well select at random from a pool of all available reading items of some specified difficulty level (a) a sample of all items based on passages containing more than 200 words and (b) a sample based on passages containing less than 100 words (it is assumed here that there is only one item per reading passage). He then places these items in random order and administers them to a group of examinees, obtaining separate scores for the long and for the short items. He computes the validity of each score, using some available criterion. If the two validity coefficients differ by little more than the type 2 standard error of their difference, it seems likely that the difference is attributable to chance fluctuations due to the sampling of items. If they differ by several times this standard error, the opposite conclusion may be reached; insofar as other uncontrolled experimental variables are ruled out, the difference may plausibly be attributed to length of reading passage.

A note of caution is necessary in using the type 2 standard error formulas. These formulas involve no assumptions beyond random sampling and large n ; however, it is not at present known just how large an n is needed in any given case. The formulas in Table 1, therefore, should be used with some caution. This is particularly true of the last three rows of the table, since the correlation coefficients given in the first column undoubtedly have sharply skewed distributions when n is small.

It should, finally, be noted that the assumption of random sampling of items cannot be expected to hold for speeded tests, and the formulas given in the present paper must be considered inapplicable.

Standard Errors of Measurement and Test Reliability

Table 1 gives a practical approximation to $S.E.(t_a)$ in terms of observed sample statistics; the rigorously accurate value, as shown in a later section is

$$S.E.^2(t_a) = \frac{1}{n}\tau_a(n - \tau_a) \ . \tag{2}$$

Here $\tau_a = E(t_a)$ is the true score of examinee a, i.e., the expected value* of t_a over all randomly parallel forms of the test. The standard error of the score of an examinee is the standard deviation of the errors of measurement of his score (error of measurement $= t_a - \tau_a$). The average of such standard deviations of errors of measurement over all examinees,

$$\frac{1}{N}\sum_a S.E.^2(t_a) = \frac{1}{N}\sum_a E(t_a - \tau_a)^2 \ , \tag{3}$$

*The expectation symbol, E, denotes the average (arithmetic mean) value over all type 2 samples. Thus the operator E can be treated by the same rules as a summation sign, so that $E(x + y) = E(x) + E(y)$, $E\sum_a(t_a) = \sum_a E(t_a)$, $E(n\bar{t}) = nE(\bar{t})$, $E(\tau_a) = \tau_a$, etc. By definition $\tau_a = E(t_a)$, $S.E.^2[f(t)] = E[f(t) - E\{f(t)\}]^2 = E\{f(t)\}^2 - [E\{f(t)\}]^2$, and $\mathrm{Cov}\left[f_1(t),f_2(t)\right] = E\left[f_1(t)f_2(t)\right] - E\left[f_1(t)\right]E\left[f_2(t)\right]$, where $f(t)$ is any function of t.

may appropriately be compared with the conventional "standard error of measurement" of test theory. This latter, which will be denoted by "S.E.Meas.," is likewise an average over all examinees. It is conventionally defined by the formula

$$\text{S.E.Meas.} = s_t\sqrt{1 - \text{reliability}} \quad . \tag{4}$$

Specifically, it will now be shown that the squared standard error of measurement given by equation 3 is exactly equal to that which would be expected in equation 4 if the test reliability there were given by the Kuder-Richardson formula 21 in reference (5). In our notation, this formula is

$$r_{21} = \frac{n}{n-1}\,\frac{s_t^2 - \bar{t}(1 - \bar{t}/n)}{s_t^2} \quad . \tag{5}$$

Averaging equation 2 over all examinees, we find

$$\frac{1}{N}\sum_a \text{S.E.}^2(t_a) = \frac{1}{nN}\sum_a \gamma_a(n - \gamma_a)$$

$$= \frac{1}{N}\sum_a \gamma_a - \frac{1}{nN}\sum_a \gamma_a^2$$

$$= \bar{\gamma} - \frac{1}{n}(\sigma_\gamma^2 + \bar{\gamma}^2) \quad . \tag{6}$$

From (5) and (4), the expected value of the squared S.E.Meas. is

$$E\left[s_t^2(1 - r_{21})\right] = E\left[\frac{1}{n - 1}\,(n\bar{t} - \bar{t}^2 - s_t^2)\right] \quad . \tag{7}$$

In order to deal with (7) we first need expressions for $E(s_t^2)$ and $E(\bar{t})^2$.

$$E(s_t^2) = E\left[\frac{1}{N}\sum_a (t_a - \bar{t})^2\right] = E\left[\frac{1}{N}\sum_a \left\{(t_a - \tau_a) + (\tau_a - \bar{\tau}) - (\bar{t} - \bar{\tau})\right\}^2\right] . \quad (8)$$

After squaring and rearranging E and Σ signs,

$$E(s_t^2) = \frac{1}{N}\left[\sum_a E\left\{(t_a - \tau_a)^2\right\} + E\left\{\sum_a(\tau_a - \bar{\tau})^2\right\} + NE\left\{(\bar{t} - \bar{\tau})^2\right\} + 2\sum_a(\tau_a - \bar{\tau})E\left\{(t_a - \tau_a)\right\} - 2E\left\{(\bar{t} - \bar{\tau})\sum_a(t_a - \tau_a)\right\} - 2E\left\{(\bar{t} - \bar{\tau})\sum_a(\tau_a - \bar{\tau})\right\}\right] . \quad (9)$$

Now the fourth and the last terms on the right vanish since $E(t_a - \tau_a)$ and $\sum_a(\tau_a - \bar{\tau})$ both equal zero. It is seen that we have, term for term,

$$E(s_t^2) = \frac{1}{N}\sum_a \text{Var } t_a + \sigma_\tau^2 + \text{Var } \bar{t} + 0 - 2 \text{ Var } \bar{t} - 0 \quad . \quad (10)$$

Now $\text{Var } t_a$ is given by (2), so that

$$E(s_t^2) = \frac{1}{nN}\sum_a \tau_a(n - \tau_a) + \sigma_\tau^2 - \text{Var } \bar{t} \quad . \quad (11)$$

Finally, proceeding as in (6), we have

$$E(s_t^2) = \bar{\tau} - \frac{1}{n}\bar{\tau}^2 + \frac{n-1}{n}\sigma_\tau^2 - \text{Var } \bar{t} \quad . \quad (12)$$

Next,

$$E(\bar{\bar{t}}^2) = E[(\bar{\bar{t}} - \bar{\bar{\tau}}) + \bar{\bar{\tau}}]^2$$

$$= E(\bar{\bar{t}} - \bar{\bar{\tau}})^2 + 2\bar{\bar{\tau}}E(\bar{\bar{t}} - \bar{\bar{\tau}}) + E(\bar{\bar{\tau}}^2)$$

$$= \text{Var } \bar{\bar{t}} + \bar{\bar{\tau}}^2 \quad . \qquad (13)$$

From (7), (12), and (13),

$$E\left[s_t^2(1 - r_{21})\right] = \frac{1}{n-1}\left[n\bar{\bar{\tau}} - \text{Var } \bar{\bar{t}} - \bar{\bar{\tau}}^2 - \bar{\bar{\tau}} + \frac{1}{n}\bar{\bar{\tau}}^2 - \frac{n-1}{n}\sigma_{\tau}^2 + \text{Var } \bar{\bar{t}}\right]$$

$$= \bar{\bar{\tau}} - \frac{1}{n}\sigma_{\tau}^2 - \frac{1}{n}\bar{\bar{\tau}}^2 \quad . \qquad (14)$$

This result is the same as that in (6). We have shown that the average squared standard error of measurement found in type 2 sampling is exactly equal to the expected value of the squared S.E.Meas. derived from the formula 21 Kuder-Richardson reliability coefficient.

The logical relation between Kuder-Richardson formulas 20 and 21 can be derived from equations 1 and 5, from which it is readily found that

$$s_t^2(1 - r_{20}) = s_t^2(1 - r_{21}) - \frac{n^2}{n-1}s_p^2 \quad . \qquad (15)$$

Now the term on the left and the first term on the right of (15) are the squared standard errors of measurement computed from r_{20} and from r_{21} , respectively. Furthermore, since $ns_p^2/(n - 1)$ is the best unbiassed small-sample estimate of the population variance σ_p^2, it is seen that the last term on the right is the small-sample estimator for the squared standard error of the mean score (see equation 22). Consequently, we may rewrite (15) as

$$(\text{S.E.Meas.}_{20})^2 = (\text{S.E.Meas.}_{21})^2 - \text{S.E.}^2(\bar{t}) \;. \qquad (16)$$

The difference between r_{20} and r_{21}, as made apparent in equation 16, arises from the fact that some randomly parallel forms are, by chance, composed of harder-than-average items, or of easier-than-average items; consequently, the mean of the actual scores on any given test is not exactly equal to the mean of the true scores for the same examinees. The use of r_{20} is appropriate whenever one is willing to ignore any difference between the mean test score of the group and their mean true score, i.e., when one is concerned only with the relative rather than the absolute size of the scores of the group. On the other hand, r_{21} should be used whenever one is concerned with the actual magnitude of the errors of measurement, e.g., whenever there is a predetermined cutting score which divides the examinees into passing and failing groups.

Comparison with Certain Standard Formulas

A formula closely related to equation 4 is the following (adapted from equation 66 of reference (7)):

$$\mathrm{S.E.}(\bar{t}) = \frac{s_t}{\sqrt{N}} \sqrt{1 - \mathrm{reliability}} \quad . \tag{66}$$

The question arises as to why S.E.$(\bar{t})$ in equation 66 has a totally different formula from that given in Table 1 for the type 2 standard error of the mean. If we use equation 66 to determine whether or not two forms of a test yield significantly different mean scores, we will always find the difference to be significant provided only that we take a sufficiently large number of examinees (N) for our experiment. This is true because the standard error of equation 66 is inversely proportional to $\sqrt{N}$ -- the standard error vanishes when N is large. In spite of this fact, it should be noted that (66) is not a type 1 standard error. A type 1 standard error involves the sampling of individuals, whereas only a single group of examinees is contemplated in (66).

The standard error given in equation 66 represents only the sampling fluctuation due to those errors of measurement that "average out" when taken over many individuals. Such errors of measurement arise from virtually instantaneous "chance" fluctuations in the individual. One example of such an error of measurement is the following: An examinee, not knowing the answer to a true-false item tosses a coin, in effect, to select the correct answer. If the same test could be administered again without practice effect, the same examinee would have a fifty-fifty chance of giving a different answer. This difference gives rise to an error of measurement of the type under discussion.

The standard error of the mean given in Table 1 includes not only sampling errors of the sort just mentioned, but also sampling errors arising from the sampling of the test items.

The line of reasoning applied to equation 66 is equally applicable to Wilks' (10) and to Votaw's (9) significance tests when either of these is used as a criterion of "parallelism" in tests, as suggested by Gulliksen (3, Ch. 14). Gulliksen defines "parallel" tests as having equal means, equal variances, and equal intercorrelations with each other and with all external criteria (as well as satisfying appropriate non-statistical criteria of parallelism). Wilks' and Votaw's significance tests provide rigorous statistical criteria for "parallelism" under this definition. It would not be very desirable, however, to apply Wilks' or Votaw's procedures to data such as were obtained in the second illustrative example given in a preceding section. If a test composed of items having a certain characteristic is to be compared with a test composed of different items having a second characteristic, it may not be very useful to set up the null hypothesis that the two tests are strictly interchangeable in every way. Such a null hypothesis will always be rejected if N is sufficiently large, but the rejection of this hypothesis does not necessarily imply that the first and second characteristics have different effect, since the observed discrepancy might be readily accounted for as no greater than would be expected to be found in comparing two randomly parallel tests composed of the same kind of items.

Sampling Distributions of Test Statistics

It remains only to present the derivations of the results that have up to now been quoted without proof. The derivations are based on the assertion that there is a definite response (x_{ia}) that a given examinee will make to a given item. The nature of this response may or may not be known in advance. The group of N examinees to whom the items or tests are administered is a fixed group not subject to sampling fluctuation or other changes.

The responses of the N examinees to item i may be specified by the column vector $\{x_i = x_{i1}, x_{i2}, \ldots, x_{iN}\}$. Since each item response is assumed to be treated as either "right" or "wrong", $x_{ia} = 0$ or 1 , and there are exactly 2^N possible different vectors, i.e., different patterns of item response. If we let the subscript $I = 1,2,3,\ldots,2^N$, then these possible patterns are represented by the 2^N vectors x_I . If two items have exactly the same pattern of responses, i.e., if the response of each examinee is the same on both items, then the two items are wholly indistinguishable in the present situation. It may therefore be asserted without loss of generality that, for present purposes, any infinite pool of items is composed of 2^N different kinds of items, designated by the 2^N vectors x_I . The relative frequencies of occurrence of the different kinds of items are therefore the only parameters needed to describe completely any infinite pool; these parameters will be denoted by π_I , the relative frequencies of occurrence of the patterns x_I .

When a random sample of n test items is drawn from the pool, the probability that the resulting n - item test will be composed of n_1 items of the first kind, n_2 items of the second kind, ..., n_I items of the I - the kind, ..., $n_{(2^N)}$ items of the 2^N - th kind is given by the standard multinomial distribution (6, pp. 58-59):

$$f(n_1, n_2, \ldots, n_{(2^N)}) = \frac{n!}{\prod_I n_I!} \prod_I \pi_I^{n_I} \quad . \tag{17}$$

It can be shown (1, p. 419) that the quantities $V_I = (n_I - n\pi_I)/\sqrt{n\pi_I}$ are asymptotically normally distributed for large n with zero means and with the (singular) variance-covariance matrix $I - \pi\pi'$, where I is the identity matrix and π is the column vector $(\sqrt{\pi_1}, \sqrt{\pi_2}, \ldots, \sqrt{\pi_{(2^N)}})$ Now, the test score of individual a is $z_a = \frac{1}{n}\sum_i x_{ia} = \frac{1}{n}\sum_I x_{Ia} n_I$, the x_{Ia} being given constants, 0 or 1, not subject to sampling fluctuation; or, in terms of V_I , $z_a = \sum_I \pi_I x_{Ia} + \frac{1}{\sqrt{n}} \sum_I \sqrt{\pi_I} x_{Ia} V_I$. The first term on the right is $\zeta_a = \mathcal{T}_a/n$, the "true" proportion-correct score; so that, finally, $\sqrt{n}(z_a - \zeta_a) = \sum_I \sqrt{\pi_I} x_{Ia} V_I$. It is thus seen that the N variables $\sqrt{n}(z_a - \zeta_a)$ are asymptotically jointly multinormally distributed, each with a mean of zero, a variance which turns out to be $\zeta_a(1 - \zeta_a)$, and covariances $\zeta_{ab} - \zeta_a\zeta_b$, where ζ_{ab} is the proportion of all items answered correctly by both examinee a and examinee b . It follows immediately that the large-sample standard error of z_a is $\sqrt{\zeta_a(1 - \zeta_a)/n}$ (cf. (2)). The derivation of these and other standard errors will be left to the following section, however.

By a well-known theorem, if $f(z_1, z_2, \ldots, z_n)$ is a function of the z_a having continuous first-order partial derivatives with respect to each z_a at the point $(\zeta_1, \zeta_2, \ldots, \zeta_N)$, and if at least one of these derivatives is nonvanishing at this point, then the quantity $\sqrt{n}\left[f(z_1, z_2, \ldots, z_N) - f(\zeta_1, \zeta_2, \ldots, \zeta_N)\right]$ is asymptotically normally distributed with zero mean when n is sufficiently large. This theorem assures us that the mean score ($\bar{z}$ or $\bar{t}$), the standard deviation of the scores (s_z or s_t), the Kuder-Richardson formula 21 reliability ($r_{21\cdot}$) , and the test validity (r_{cz} or r_{ct}), are approximately normally distributed in type 2 sampling with large n ; and in addition gives us the large-sample expected value of each statistic. It seems highly likely that the Kuder-Richardson reliability, formula 20, likewise is asymptotically normally distributed, but no proof of this conclusion is available at present, in view of the fact that the formula for this statistic involves $\sigma^2(p)$, which is not a function of the z_a .

Derivations of Expected Values and Standard Errors

The Individual Score

The proportion of the items in the entire pool to which examinee a will give the correct answer is, by definition, $\zeta_a = \tau_a/n$. If n items are drawn at random from the pool, t_a , the score of examinee a on the resulting test, i.e., the number of items that he will answer successfully, will of necessity have the usual binomial distribution* with mean and variance

$$E(t_a) = \tau_a \quad , \tag{18}$$

$$S.E.^2(t_a) = \frac{1}{n} \tau_a(n - \tau_a) = n\zeta_a(1 - \zeta_a) \ . \tag{19}$$

This conclusion (and also those that follow, except as large n may be assumed) depends on no assumptions whatever except that of random sampling. Equation 19 is identical with equation 2, which was discussed in a previous section. If the observed value t_a is substituted for the unknown τ_a in (19), we obtain the square of the first formula of Table 1.

For finite sampling, when n items are drawn without replacement from a finite pool of m items, the corresponding formulas, stated without proof, are

$$E(t_a) = \tau_a \ , \tag{18*}$$

*If we concern ourselves with only a single examinee, the number of correct responses that he gives on one sample of items is not correlated with the number that he gives on other samples.

$$\text{S.E.}^2(t_a) = \frac{m-n}{mn}\tau_a(n - \tau_a) \quad . \tag{19'}$$

The Mean Score of the Group Tested

It should be noted that the scores of examinees a and b are not independent over different parallel forms of the test. If a particular form happens to be composed of rather difficult items, both examinees will tend to get low scores; if a particular form happens to be easy, both will tend to score higher. Consequently, although the expected value of the mean score in the group is equal to the mean of the expected values of the individual scores, i.e.,

$$E(\bar{t}) = \frac{1}{N_a}\Sigma\tau_a = \bar{\tau} \quad , \tag{20}$$

the standard error of the mean is not an average of the standard errors of the individual scores.

It will be convenient from this point on to work with $z_a = t_a/n$, the proportion-correct score, rather than with t_a itself. The nature of the desired standard error follows immediately from the fact that the mean score ($\bar{z}$) is identically equal to the average item difficulty

$$\bar{z} \equiv M(p) \quad . \tag{21}$$

The usual formulas for the standard error of a mean apply to $M(p)$, so that

$$\text{S.E.}^2(\bar{z}) = \frac{1}{n}\sigma^2(p) \tag{22}$$

where $\sigma(p)$ is the standard deviation of the item difficulties over the whole pool of items.* If the observed value of $s^2(p)$ is substituted for the unknown $\sigma^2(p)$, and if t/n is substituted for z, the square of the second formula of Table 1 is obtained.

In sampling from a finite pool of m items, the corresponding formula, stated without proof, is

$$S.E.^2(\bar{z}) = \frac{m-n}{mn}\sigma^2(p) \quad . \tag{22'}$$

We may note that $\sigma(p)$ for a given set of items, and hence $S.E._2(\bar{z})$ for a given test, will be higher when N is small than when N is large. Suppose, for example, that all items have the same difficulty (p) for a very large group of examinees, so that for this group $\sigma(p) = 0$. If the same items are administered to a smaller group of examinees drawn at random from the larger, the observed values of p_i in the smaller group will differ from each other because of type 1 sampling fluctuations, and $\sigma(p)$ will be greater than zero. In the extreme case where $N = 1$, the observed values of p are of necessity either 0 or 1, and $\sigma(p)$ is at a maximum.

*Equation 19 is a special case of equation 22, being obtained when $p_i = x_{ia}$.

The Standard Deviation of the Scores of the Group Tested

In order to obtain the standard error of s_z^2 , we first use the formula for the variance of a sum to write

$$s_z^2 = \frac{1}{n^2}\sum_{hi}\sum s_{ih} \tag{23}$$

s_{ih} being the covariance between item i and item h . Then, again from the formula for the variance of a sum,

$$\text{Var } s_z^2 = \frac{1}{n^4}\sum_{hijk}\sum\sum\sum \text{Cov}(s_{ih},s_{jk}) \quad , \tag{24}$$

where "Cov" stands for the sampling covariance: $\text{Cov}(s_{ih},s_{jk}) = Es_{ih}s_{jk} - Es_{ih}Es_{jk}$.

Grouping the sums in (24), we obtain

$$\text{Var } s_z^2 = \frac{1}{n^4}\Bigg[\overset{n^4-6n^3+11n^2-6n}{\underset{(h,\ i,\ j,\ k\ \neq)}{\Sigma\quad\Sigma\quad\Sigma\quad\Sigma}} \text{Cov}(s_{hi},s_{jk}) + 2\overset{n^3-3n^2+2n}{\underset{(i,j,k\ \neq)}{\Sigma\quad\Sigma\quad\Sigma}} \text{Cov}(s_i^2,s_{jk})$$

$$+ 4\overset{n^3-3n^2+2n}{\underset{(i,j,k\ \neq)}{\Sigma\quad\Sigma\quad\Sigma}} \text{Cov}(s_{ij},s_{jk}) + 4\overset{n^2-n}{\underset{i\neq j}{\Sigma\quad\Sigma}} \text{Cov}(s_i^2,s_{ij}) + \text{other sums}$$

$$\text{containing no more than } n^2 \text{ terms each}\Bigg] . \tag{25}$$

Here the first sum is over all sets of four subscripts no two of which are the same, etc. The coefficient 2 of the second sum arises from combining the two equivalent expressions $\underset{(i,j,k\neq)}{\Sigma\,\Sigma\,\Sigma}\text{Cov}(s_i^2,s_{jk})$ and

$\Sigma\,\Sigma\,\Sigma_{(h,i,j\neq)}\ \mathrm{Cov}(s_{hi}, s_j^2)$. The other numerical coefficients arise similarly. The polynomials in n written above the summation signs indicate the number of terms involved in the summation.

Now, the terms under each summation sign in (25) are all the same no matter what the numerical values of the subscripts; consequently

$$\mathrm{Var}\ s_z^2 = \frac{1}{n^4}\Big[(n^4 - 6n^3 + 11n^2 - 6n)\mathrm{Cov}(s_{hi}, s_{jk}) + 2(n^3 - 3n^2 + 2n)\mathrm{Cov}(s_i^2, s_{jk})$$

$$+ 4(n^3 - 3n^2 + 2n)\mathrm{Cov}(s_{ij}, s_{jk}) + O(n^2)\Big] , \qquad (26)$$

where $O(n^2)$ stands for terms of order n^2 . In (26) and in the following paragraph it is understood that $h,i,j,k\neq$.

Now, s_{hi} and s_{jk} fluctuate independently over successive samples, so that $\mathrm{Cov}(s_{hi}, s_{jk}) = 0$. The same is true of s_i^2 and s_{jk} . Consequently,

$$\mathrm{Var}\ s_z^2 = \frac{4}{n^4}(n^3 - 3n^2 + 2n)\mathrm{Cov}(s_{ij}, s_{jk}) + O\left(\frac{1}{n^2}\right) = \frac{4}{n}\mathrm{Cov}(s_{ij}, s_{jk}) + O\left(\frac{1}{n^2}\right) . \qquad (27)$$

Equation 27 gives the desired result, but not in a very useful form, since $\mathrm{Cov}(s_{ij}, s_{jk})$ is a function of population parameters and is generally not known. As a final step, then, it will be shown that $s^2(s_{iz})$, the actual variance (over items 1 to n) of the observed item-test covariances, provides a "consistent" estimate of $\mathrm{Cov}(s_{ij}, s_{jk})$, i.e., it will be proved that

$$Es^2(s_{iz}) = \mathrm{Cov}(s_{ij}, s_{jk}) + O\left(\frac{1}{n}\right) \; . \tag{28}$$

From the formula for the covariance of a sum,

$$s_{iz} = \frac{1}{n}\sum_j s_{ij} \; ; \tag{29}$$

$$s^2(s_{iz}) = \frac{1}{n^2}\sum_{jk}\sum s(s_{ij}, s_{ik}) \; , \tag{30}$$

the term under the summation sign being the actual covariance (over items 1 to n) of the observed values of s_{ij} and s_{ik} :

$$s(s_{ij}, s_{ik}) = \frac{1}{n}\sum_i s_{ij}s_{ik} - \frac{1}{n^2}\left(\sum_i s_{ij}\right)\left(\sum_i s_{ik}\right) \; . \tag{31}$$

Substituting from (31) into (30), and taking expected values, we find

$$Es^2(s_{iz}) = \frac{1}{n^3}\sum_{ijk}\sum\sum Es_{ij}s_{ik} - \frac{1}{n^4}\sum_{hijk}\sum\sum\sum Es_{hj}s_{ik} \; . \tag{32}$$

Grouping the sums on the right, we have

$$Es^2(s_{iz}) = \frac{1}{n^3}\left[\overset{n(n-1)(n-2)}{\underset{(i,\ j,\ k\neq\ )}{\Sigma\ \ \Sigma\ \ \Sigma}} Es_{ij}s_{ik} + O(n^2)\right] - \frac{1}{n^4}\left[\overset{n(n-1)(n-2)(n-3)}{\underset{(h,\ i,\ j,\ k\neq\ )}{\Sigma\ \ \Sigma\ \ \Sigma\ \ \Sigma}} Es_{hj}s_{ik}\right.$$

$$\left. + \; O(n^3)\right] \; . \tag{33}$$

Now, the terms under each summation sign in (33) are the same regardless of the numerical value of the subscript. Furthermore, as already pointed out in deriving (27), $\mathrm{Cov}(s_{hi}, s_{jk}) = 0$ when $h, i, j, k \neq$, or in other words, $Es_{hj}s_{ik} - Es_{hj}Es_{ik} = 0$, or $Es_{hj}s_{ik} = Es_{ij}Es_{ik}$. Consequently,

$$Es^2(s_{iz}) = Es_{ij}s_{ik} - Es_{ij}Es_{ik} + O\left(\frac{1}{n}\right) . \qquad (34)$$

But this is the same as (28), which was to be proved.

The large sample standard error of s_z^2 may therefore be estimated from the actual variance of the observed item-test covariances:

$$S.E.^2(s_z^2) = \frac{4}{n} s^2(s_{iz}) . \qquad (35)$$

By means of the "delta" method (4, Vol. 1, pp. 208 ff.), it is readily shown from (35) that in large samples

$$S.E.^2(s_z) = \frac{1}{4s_z^2} S.E.^2(s_z^2) = \frac{s^2(s_{iz})}{ns_z^2} . \qquad (36)$$

If t/n is substituted for z in (36), the square of the third equation of Table 1 is obtained.

The corresponding squared standard error for sampling from finite populations may be shown to be

$$S.E.^2(s_z^2) = 4 \frac{m-n}{mn} s^2(s_{iz}) . \qquad (37)$$

The Kuder-Richardson Reliability Coefficient, Formula 20

Let the usual formula for r_{20} , the Kuder-Richardson formula 20, be rewritten as follows:

$$r_{20} = \frac{n}{n-1}\left(1 - \frac{R}{n}\right) , \tag{38}$$

where $R = \frac{1}{n}\Sigma_i s_i^2/s_z^2 = M/s_z^2$, say .

In the extraordinary case where $s_z^2 = 0$, we will agree not to try to compute any value of r_{20} . The "delta" method may now be used to obtain the result.

$$\text{Var } R \doteq \frac{1}{s_z^4} \text{ Var } M + \frac{M^2}{s_z^8} \text{ Var } s_z^2 - \frac{2M}{s_z^6} \text{ Cov}(M, s_z^2) . \tag{39}$$

Now $\text{Var}(s_z^2)$ is already known from equation 35. $\text{Var}(M)$ can be evaluated by the usual formula for the standard error of a mean:

$$\text{Var } M \doteq \frac{1}{n} s^2(s_i^2) , \tag{40}$$

where $s^2(s_i^2)$ is the actual variance of the observed item variances. Finally, it is readily shown, by methods similar to those used in evaluating $\text{Var }(s_z^2)$, that

$$\text{Cov}(M, s_z^2 \doteq \frac{2}{n} s(s_i^2, s_{iz}) , \tag{41}$$

where $s(s_i^2, s_{iz})$ is the actual covariance between the observed item variances and the observed item-test covariances.

Consequently,

$$\mathrm{Var}\ R \doteq \frac{1}{n s_z^4}\left[s^2(s_i^2) + 4R^2 s^2(s_{iz}) - 4R s(s_i^2, s_{iz})\right] \quad . \qquad (42)$$

Now $\mathrm{Var}(r_{20}) = \frac{1}{n^2}\mathrm{Var}(R)$; hence, to order $1/n^4$,

$$\mathrm{S.E.}^2(r_{20}) = \frac{1}{n^3 s_z^4}\left[s^2(s_i^2) + 4n^2(1 - r_{20})^2 s^2(s_{iz}) - 4n(1 - r_{20}) s(s_i^2, s_{iz})\right] . \qquad (43)$$

It may be noted that the quantity $(1 - r_{20})$ is of order $1/n$, because, by the Spearman-Brown formula, $\lim_{n=\infty} n(1 - r_{20}) = \text{constant}$. It is then seen from (43) that $\mathrm{S.E.}^2(r_{20})$ is a quantity of order $1/n^3$. Equation 43 leads directly to the fourth formula of Table 1.

It may be shown that the corresponding standard error when sampling from a finite population is $(m - n)/m$ times the value given in (43).

The Kuder-Richardson Reliability Coefficient, Formula 21

By a procedure wholly parallel to that used for the formula-20 reliability coefficient, it is found that, approximately,

$$\mathrm{S.E.}^2(r_{21}) = \frac{1}{n^3 s_z^4}\Big[(1 - 2\bar{z})^2 s^2(p) + 4n^2(1 - r_{21})^2 s^2(s_{iz})$$

$$- 4n(1 - r_{21})(1 - 2\bar{z}) s(p_i, s_{iz})\Big] \quad , \qquad (44)$$

where $s(p_i, s_{iz})$ is the actual covariance between the observed item difficulties and the observed item-test covariances. Equation 44 leads directly to the fifth formula of Table 1.

The standard error of the split-half reliability coefficient has not been worked out. It must, however, be larger than the standard error of r_{20}, given by (43), since r_{20} is the mean of the split-half coefficients from all possible splits, as shown by Cronbach (2).

The Validity Coefficient

If c is an outside criterion,

$$r_{cz} = \frac{s_{cz}}{s_c s_z} \quad . \tag{45}$$

By the "delta" method,

$$\text{Var } r_{cz} = r_{cz}^2 \left[\frac{\text{Var } s_{cz}}{s_{cz}^2} + \frac{\text{Var } s_z^2}{4s_z^4} - \frac{\text{Cov}(s_{cz}, s_z^2)}{s_{cz} s_z^2} \right] \tag{46}$$

It is found that

$$\text{Var } s_{cz} \doteq \frac{1}{n} s^2(s_{ci}) \quad ; \tag{47}$$

$$\text{Cov}(s_{iz}, s_z^2) \doteq \frac{2}{n} s(s_{ci}, s_{iz}) \quad . \tag{48}$$

Finally,

$$\text{S.E.}^2(r_{cz}) \doteq \frac{1}{n s_z^2} \left[\frac{1}{s_c^2} s^2(s_{ci}) - \frac{2r_{cz}}{s_z s_c} s(s_{ic}, s_{iz}) + \frac{r_{cz}^2}{s_z^2} s^2(s_{iz}) \right] \quad . \tag{49}$$

Equation 49 leads directly to the last formula of Table 1.

The corresponding standard error for sampling from a finite pool of items is presumably $(m - n)/m$ times the foregoing quantity.

References

1. Cramer, H. Mathematical methods of statistics. Princeton Univ. Press, 1946.

2. Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, *16*, 297-334.

3. Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

4. Kendall, M. G. The advanced theory of statistics. London: Charles Griffin and Co., 1948. 2 vols.

5. Kuder, G. F. and Richardson, M. W. The theory of the estimation of test reliability. *Psychometrika*, 1937, *2*, 151-160.

6. Mood, A. M. Introduction to the theory of statistics. New York: McGraw-Hill, 1950.

7. Peters, C. C. and Van Voorhis, W. R. Statistical procedures and their mathematical bases. New York: McGraw-Hill, 1940.

8. Tucker, L. R. A note on the estimation of test reliability by the Kuder-Richardson formula (20). *Psychometrika*, 1949, *14*, 117-119.

9. Votaw, D. F., Jr. Testing compound symmetry in a normal multivariate distribution. *Ann. math. Statist.*, 1948, *19*, 447-473.

10. Wilks, S. S. Sample criteria for testing equality of means, equality of variances, and equality of covariances in a normal multivariate distribution. *Ann. math. Statist.*, 1946, *17*, 257-281.